WordNet relációk szerepének vizsgálata a jelentés-egyértelműsítésben

Szarvas György[1], Csendes Dóra[1] és Kocsor András[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
6720 Szeged, Árpád tér 2., Hungary
{dcsendes, szarvas}@inf.u-szeged.hu
[2] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport,
6720 Szeged, Aradi vértanúk tere 1., Hungary
kocsor@inf.u-szeged.hu

1 Bevezetés

A jelentés egyértelműsítés feladata az egyes szóalakok konkrét jelentésének meghatározása, a szövegkörnyezet segítségével. A lehetséges jelentések halmaza általában rendelkezésre áll valamilyen elektronikus szótár, vagy ontológia/nyelvi adatbázis formájában. Ebben a cikkben ismertetjük az angol nyelvű WordNet [1] ontológia különböző relációinak jelentés-egyértelműsítésben való felhasználhatóságát vizsgáló kutatásunkat.

Gyakran felügyelt tanulási módszereket alkalmaznak a szóalakok adott szövegkörnyezetben legvalószínűbb jelentésének kiválasztására, melyhez kézileg egyértelműsített példákat tartalmazó korpuszra van szükség. Ilyen korpusz angol nyelvre a SemCor [1] korpusz, mely ingyenesen hozzáférhető. Kísérleteink során ezt a korpuszt fogjuk használni a paraméterek hangolására, illetve tesztelési célokra.

A másik gyakori megközelítés nem igényli előre annotált korpusz meglétét [5, 7, 8], helyette az elektronikus szótárban kódolt információ (leggyakrabban a szóalakok definíciói, glosszái, illetve a szótári egységek között definiált relációk) képezi az egyértelműsítés alapját. Ezek a módszerek a szavak glosszái közötti egyezéseket, illetve a relációs gráfban a szóalakok között értelmezhető utak hosszát, mint szemantikai értelemben vett távolságot használják fel.

Természetesen találhatóak a sokféle heurisztikát ötvöző jelentés-egyértelműsítő rendszerek is [6].

2 Kísérletek

A folyó szövegek szavaihoz azok WordNet ontológiabeli jelentésének hozzárendelésére a mondat szavainak egymástól vett ontológiabeli távolságát vizsgáltuk. Módszerünk mögött az a feltevés húzódik meg, hogy a mondatok szavai szemantikailag koherens struktúrát alkotnak az ontológia gráfjában, azaz a jelentés-hozzárendelés elvé-

gezhető a szóalakok valamely távolságmetrika mellett vett legközelebbi rendszerének megkeresésével.

Számos szemantikai hasonlóságot gráfbeli távolsággal definiáló módszer létezik, azonban ezek legtöbbször az ontológia hierarchikus (hipo-, hipernim) relációit veszik figyelembe [3], vagy más területre – weblapok rendszerezett adatbázisán – lettek kifejlesztve [4]. Munkánk célja a sokféle WordNet-reláció fontosságának kivizsgálása a jelentés azonosításban való hasznosságuk szerint, hogy a megfelelő súlyozással egy, az eddigieknél hatékonyabb, gráfbeli távolságalapú egyértelműsítő heurisztikát kapjunk (esetleg azonosítsuk az egyértelműsítés szempontjából haszontalan relációkat).

Kísérleteinkhez a WordNet ontológiát, kiértékelésre a jelentés-egyértelműsített SemCor korpuszt használtuk, a megfelelő magyar erőforrások hiányában. A kidolgozott módszerek magyar WordNet birtokában magyar nyelvre is átültethetők.

Az általunk bemutatott módszer a Bevezetésben tárgyalt két típus között helyezkedik el, hiszen az egyértelműsítéshez használt metrika az utóbbi, felügyelet nélküli módszerekkel rokon, azonban célunk volt az eddig az egyértelműsítés során kiaknázatlan relációfajták felhasználása, illetve vizsgálata a hatékonyságra gyakorolt hatásukat illetően, melyhez címkézett adatokat használhatunk validációs célokra. A szerzők tudomása szerint nem készült az összes, WordNet ontológiában tárolt relációt felhasználó távolságalapú jelentés egyértelműsítő módszer. Az egyetlen, a hierarchikus relációkon kívül mást is felhasználó eljárás Hirst és St. Onge módszere [2], valamint annak különböző változatai.

Az elvégzett kísérletek során a mondat szavainak összes jelentését figyelembe véve kiszámítjuk a szavak által kifeszített részgráfban a csúcsok távolságának összegét. A legkisebb távolságösszeggel rendelkező részgráf kijelöli a mondat szavainak az adott kontextusban legvalószínűbb jelentését. A távolság számításához az egyes relációkat különböző súlyokkal vehetjük figyelembe, ekkor minden lehetséges súlyozás egy-egy jelentés-egyértelműsítő heurisztikát definiál. A súlyok hangolásával kereshetjük a többértelműségek feloldásához legalkalmasabb súlyozást.

A lentebb található ábrán a „NOR COULD HE CALL UP (#3) MEMORY-PICTURES (#1) OF CLOSE (#2) FRIENDS (#1) OR RELATIVES (#1).” angol mondat szavai által definiált szemantikustávolság-mátrix látható. A mátrix egyes soraiban az egyértelműsítendő szavak egyes lehetséges jelentéseinek a többi szó különböző alternatíváitól való távolságait mutatja. A cél minden szóhoz egy lehetséges jelentés hozzárendelése úgy, hogy a kijelölt sorok, és oszlopok metszeteiben álló elemek összege legyen minimális. A fenti példa mondat esetén a legegyszerűbb, minden relációt egyforma, egységnyi súllyal figyelembe vevő heurisztika egyértelmű optimumot definiál, mely pontosan a jelentések helyes hozzárendelését adja.

Jelenleg is folyó kutatásunkban a következő feladatokra koncentrálunk:

- Az ábrán látható típusú mátrixokban a legkisebb össztávolságot definiáló hozzárendelést megadó hatékony algoritmus kifejlesztése
- Egyértelmű (egyetlen lehetséges jelentéssel bíró) szavak kitüntetett szerepének vizsgálata – javíthat-e a pontosságon vagy sebességen az ilyen szavak kitüntetett kezelése

- A minimális össztávolságú struktúra meghatározásakor a mondat szavainak az összes többitől vett távolságát célszerű vizsgálni, vagy elegendő egy bizonyos sugarú környezetben (a mondat távoli részei közt is mutatható ki szemantikai értelemben vett kohézió, vagy csak az egymáshoz közeli szavak között)
- Szintaktikai elemzés információinak felhasználása a keresés szűkítésére milyen hatással van az egyértelműsítésre (a keresést nem a szó szoros környezetére, hanem a vele nyelvtani kapcsolatban álló szavakra korlátozzuk)
- Változó élsúlyok használata a távolság számításakor (pl. egy hipo-hiperním kapcsolatot leíró él nem azonos szemantikai távolságot definiál a hierarchia alsóbb és felsőbb, absztraktabb fogalmakat leíró részeiben)

	call up # 1	*call up # 2*	call up # 3	*call up # 4*	memory picture # 1	*close # 1*	close # 2	*close # 3*	*close # 4*	*close # 5*	friend # 1	*friend # 2*	*friend # 3*	*friend # 4*	*friend # 5*	relative # 1	*relative # 2*
call up # 1					8	9	8	9	11	9	6	8	6	7	8	6	5
call up # 2					9	8	7	8	10	8	5	6	5	6	7	5	6
call up # 3					**6**	5	**4**	5	8	5	**5**	7	5	5	7	**5**	6
call up # 4					9	8	8	8	7	8	6	6	6	7	8	6	6
memory picture # 1	8	9	**6**	9		10	**9**	10	12	10	**7**	9	7	7	9	**7**	8
close # 1	9	8	5	8	10						6	8	6	7	8	6	7
close # 2	8	7	**4**	8	**9**						**5**	7	5	6	7	**5**	6
close # 3	9	8	5	8	10						6	8	6	7	8	6	7
close # 4	11	10	8	7	12						7	9	7	8	9	7	8
close # 5	9	8	5	8	10						6	8	6	7	8	6	7
friend # 1	6	5	**5**	6	**7**	6	**5**	6	7	6						**2**	3
friend # 2	8	6	7	6	9	8	7	8	9	8						4	5
friend # 3	6	5	5	6	7	6	5	6	7	6						2	3
friend # 4	7	6	5	7	7	7	6	7	8	7						3	4
friend # 5	8	7	7	8	9	8	7	8	9	8						4	5
relative # 1	6	5	**5**	6	**7**	6	**5**	6	7	6	**2**	4	2	3	4		
relative # 2	5	6	6	7	8	7	6	7	8	7	3	5	3	4	5		

1. ábra: A „Nor could he call up memory-pictures of close friends or relatives.” mondat szavainak szemantikustávolság-mátrixa. A nem megfelelő jelentések dőlt betűvel, a megfelelő jelentések szemantikus távolságai vastag, nagyobb számokkal szerepelnek.

3 Összegzés

Az általunk ismertetett szóalakok jelentésének WordNet-synsetek szerinti egyértelműsítése évtizedes múltra visszatekintő kutatási irányzat. Az irodalomban fellelhető egyértelműsítő eljárások egyik nagy csoportját a WordNet struktúrán, mint

címkézett, irányított gráfon értelmezett szemantikai távolság/hasonlóság metrikán alapuló algoritmusok adják, a dolgozatban ismertetett módszer is ezek közé sorolható.

A szerzők által bemutatott jelentés-egyértelműsítő az eddig ismerteken túlmutat abban, hogy a WordNet összes relációját felhasználja a gráfbeli utak hosszának vizsgálatakor, hiszen az összes ontológiai reláció szemantikai értelemben vett kapcsolatot teremt az egyes fogalmak között, így figyelembevételük a szemantikai távolság számításakor indokolt. A vizsgálatok eredményeként nemcsak egy új jelentés-egyértelműsítő heurisztikát kapunk, de a távolságmérték számításához felhasznált súlyok hangolásával a WordNet relációinak értékelését is megkapjuk, mely megmondja, hogy a kérdéses relációtípus mennyire jól használható szavak többértelműségének feloldására.

Bibliográfia

1. Fellbaum, C.: WordNet: An Electronic Lexical Database. The MIT Press, USA (1998)
2. Hirst, G., St. Onge, D.: Lexical chains as representations of context for the detection and correction of malapropisms, in: C. Fellbaum (Ed.), WordNet: An electronic lexical database, pp. 305–332., MIT Press, (1998)
3. Lin, D.: An Information-Theoretic Definition of Similarity. Proceedings of the 15th International Conf. on Machine Learning, Madison, Wisconsin (1998)
4. Maguitman, A.G., Menczer, F., Roinestad, H., Vespignani, A.: Algorithmic Detection of Semantic Similarity. Proceedings of the 14th International World Wide Web Conference, Chiba, Japan (2005)
5. McCarthy, D., Koeling, R., Weeds, J.: Ranking WordNet senses automatically, Technical Report CSRP 569. University of Sussex (2004)
6. Mihalcea, R.F., Moldovan, D.I.: A Highly Accurate Bootstrapping Algorithm for Word Sense Disambiguation. International Journal on Artificial Intelligence Tools, Vol. 10, No. 1-2 (2001)
7. Patwardhan, S., Banerjee, S., Pedersen, T.: Using Measures of Semantic Relatedness for Word Sense Disambiguation. Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City (2003)
8. Pedersen, T., Banerjee, S., Patwardhan, S.: Maximizing Semantic Relatedness to Perform Word Sense Disambiguation. Research Report UMSI 2005/25 (http://www.msi.umn.edu/general/Reports/rptfiles/2005-25.pdf) University of Minnesota, Duluth (2005)